Mean squared error estimation

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

https://cims.nyu.edu/~cfgranda/pages/MTDS_spring20/index.html

Carlos Fernandez-Granda

Prerequisites

Calculus (gradients, Hessian)

Linear algebra (vectors, matrices)

Probability (expectation, covariance matrix)

Goal

Analyze regression and linear regression from a probabilistic perspective

Constant estimate

Goal: Estimate a quantity represented by a random variable $\tilde{y}$

If we have no data (but we know the distribution) what is the best estimate in terms of mean squared error (MSE)?

$$\begin{aligned} \arg\min_{c \in \mathbb{R}} \mathrm{E}\left((c - \tilde{y})^2\right) &= \mathrm{E}\left(c^2 - 2c\tilde{y} + \tilde{y}^2\right) \\ &= c^2 - 2c\mathrm{E}\left(\tilde{y}\right) + \mathrm{E}\left(\tilde{y}^2\right) = g(c) \end{aligned}$$

Constant estimate

$$
\begin{aligned}
g(c) &:= c^2 - 2c\mathrm{E}\left(\tilde{y}\right) + \mathrm{E}\left(\tilde{y}^2\right) \\
g'(c) &= 2(c - \mathrm{E}(\tilde{y})) \\
g''(c) &= 2
\end{aligned}
$$

Convex with minimum at $\mathrm{E}(\tilde{y})$!

The mean is the best constant estimate in terms of MSE

Regression

Goal: Estimate response (or dependent variable)

Data: Several observed variables, known as features (or covariates, or independent variables)

Probabilistic perspective

Response: random variable $\tilde{y}$

Features: random vector $\tilde{x}$

What estimator (function of $\tilde{x}$) minimizes mean squared error?

Minimum mean squared error

We observe $\tilde{x} = x$

Uncertainty about $\tilde{y}$ is captured by pdf (or pmf) $f_{\tilde{y} \,|\, \tilde{x}=x}$ of $\tilde{y}$ given $\tilde{x} = x$

Let $\tilde{w}$ have that distribution

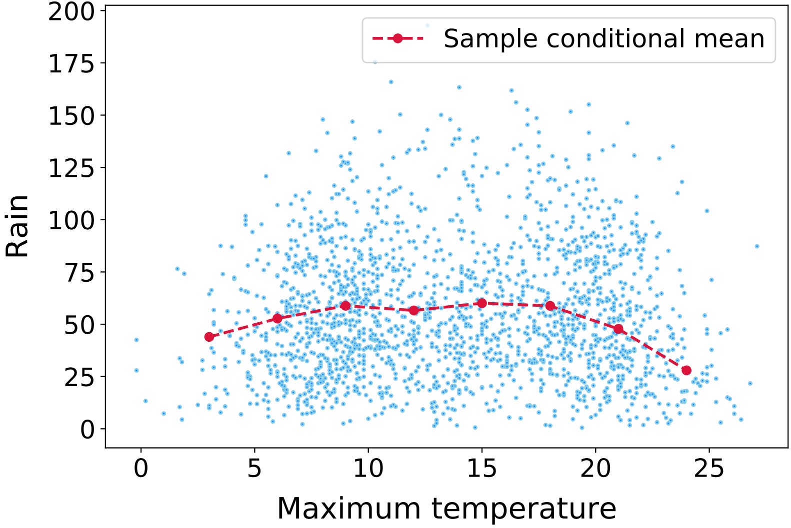
What is the minimum MSE estimate?

$$\min_{c} \mathrm{E}[(\tilde{w} - c)^2]$$

The mean of w, which equals the conditional mean

$$\mathrm{E}(\tilde{y} \,|\, \tilde{x} = x) = \int_{y \in \mathbb{R}} y \, f_{\tilde{y} \,|\, \tilde{x}} (y \,|\, x) \, \mathrm{d}x$$

Estimating rain from temperature



Are we done?

Assume we have 5 features with 100 possible values each

How many conditional averages do we need to estimate? 10^{10}!

This is known as the curse of dimensionality

Linear regression

We need to make assumptions

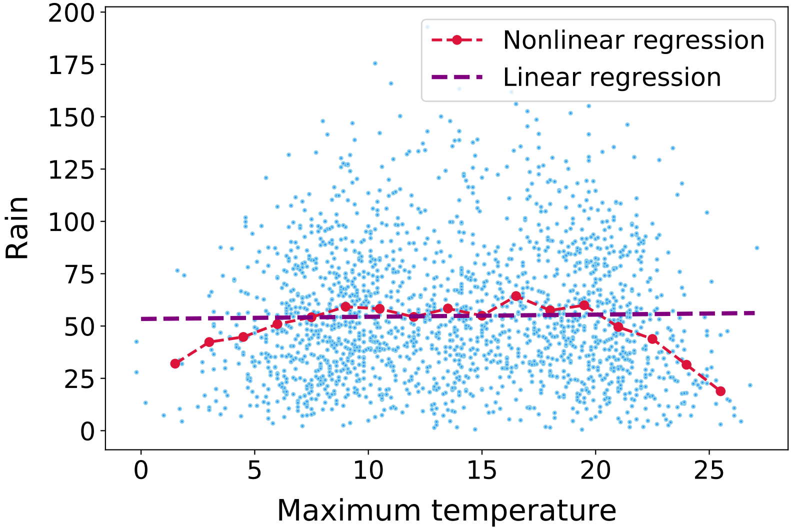
Simple but powerful assumption: Relationship is linear (or rather affine)

$$\tilde{y} \approx \beta^T \tilde{x} + \beta_0$$

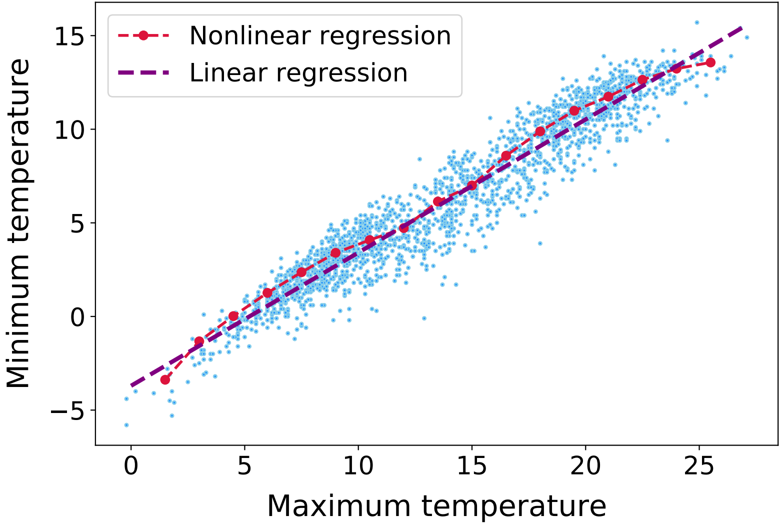
For fixed $\beta \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$

Mathematically, gradient of the regression function is constant

Estimating rain from temperature



Estimating minimum from maximum temperature



Linear regression

Constant term is a bit annoying

$$\tilde{y} \approx \beta^T \tilde{x} + \beta_0$$

Idea: Since β_0 is a constant offset, can we just center everything?

$$c(\tilde{y}) := \tilde{y} - \mathrm{E}(\tilde{y})$$

$$c(\tilde{x}) := \tilde{x} - \mathrm{E}(\tilde{x})$$

Centering

For fixed $\beta \in \mathbb{R}^p$ what is the optimal β_0?

$$\arg\min_{\beta_0} \mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta - \beta_0)^2\right] = \mathrm{E}(\tilde{y} - \tilde{x}^T\beta)$$

Plugging in:

$$\begin{aligned}\min_{\beta_0} \mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta - \beta_0)^2\right] &= \mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta - \mathrm{E}(\tilde{y}) + \mathrm{E}(\tilde{x})^T\beta)^2\right] \\ &= \mathrm{E}\left[(c(\tilde{y}) - \beta^T c(\tilde{x}))^2\right]\end{aligned}$$

From now on, everything will be centered (i.e. zero mean)

MSE

Goal: Find β minimizing

$$\mathrm{E}\left[(\tilde{y} - \tilde{x}^T\beta)^2\right] = \mathrm{E}\left(\tilde{y}^2\right) - 2\mathrm{E}\left(\tilde{y}\tilde{x}\right)^T\beta + \beta^T\mathrm{E}(\tilde{x}\tilde{x}^T)\beta$$
$$= \beta^T\Sigma_{\tilde{x}}\beta - 2\Sigma_{\tilde{y}\tilde{x}}^T\beta + \mathrm{Var}\left(\tilde{y}\right) = f(\beta)$$

where the cross-covariance vector equals

$$\Sigma_{\tilde{y}\tilde{x}}[i] := \mathrm{E}\left(\tilde{y}\,\tilde{x}[i]\right), \quad 1 \leq i \leq p$$

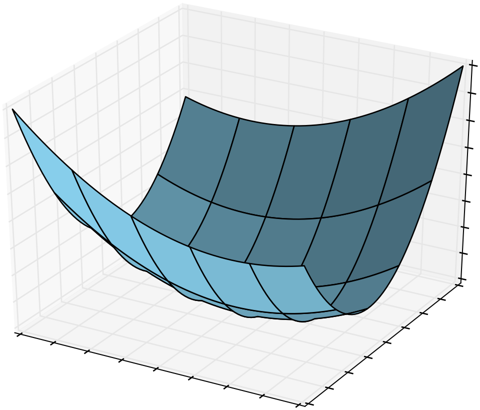
MSE function

Quadratic form

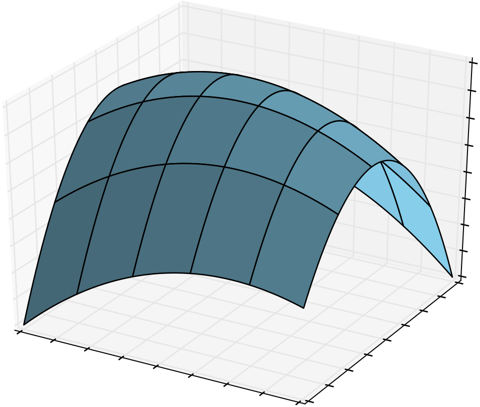
$$
\begin{aligned}
f(\beta) &:= \beta^T \Sigma_{\tilde{x}} \beta - 2\Sigma_{\tilde{y}\tilde{x}}^T \beta + \mathrm{Var}(\tilde{y}) \\
&= \beta^T A \beta + b^T \beta + c
\end{aligned}
$$

How does it look like?

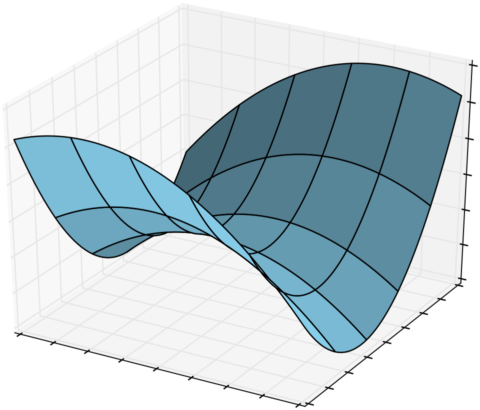
Convex?



Concave?



Neither?



Gradient and Hessian

Quadratic form

$$f(\beta) := \beta^T A\beta + b^T\beta + c$$

$$\nabla f(\beta) = 2A\beta + b$$

$$\nabla^2 f(\beta) = 2A$$

Gradient

Determines tangent plane

If gradient is zero, tangent plane is horizontal

We focus on point β^* where gradient is zero

$$\nabla f(\beta) = 2A\beta^* + b = 0$$

and rewrite the quadratic form setting

$$b = -2A\beta^*$$

Note that we have

$$\begin{aligned} f(\beta^*) &= (\beta^*)^T A\beta^* + b^T\beta^* + c \\ &= -(\beta^*)^T A\beta^* + c \end{aligned}$$

Linear minimum MSE estimator

Quadratic form

$$\begin{aligned} f(\beta) &:= \beta^T A\beta - b^T\beta + c \\ &= \beta^T A\beta - 2(\beta^*)^T A\beta + c \\ &= (\beta-\beta^*)^T A(\beta-\beta^*) - (\beta^*)^T A\beta^* + c \\ &= (\beta-\beta^*)^T A(\beta-\beta^*) + f(\beta^*) \end{aligned}$$

(assuming A is symmetric)

If for any nonzero v $v^T Av > 0$ then β^* is the solution!

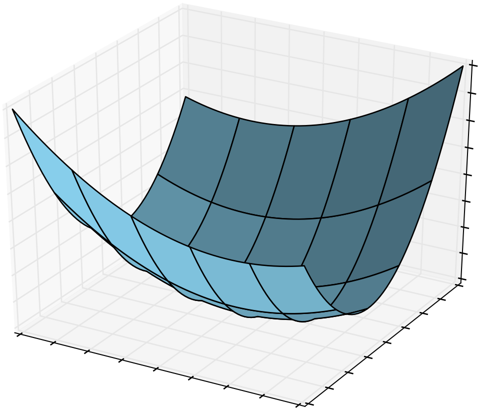
Covariance matrices are positive semidefinite

For any vector $v \in \mathbb{R}^p$

$$v^T \Sigma_{\tilde{x}} v = \mathrm{Var}\left(v^T \tilde{x}\right) \geq 0$$

If $\Sigma_{\tilde{x}}$ is full rank, then positive definite

So the MSE looks like this!



Linear minimum MSE estimator

Quadratic form

$$f(\beta) := \beta^T \Sigma_{\tilde{x}} \beta - 2\Sigma_{\tilde{y}\tilde{x}}^T \beta + \mathrm{Var}(\tilde{y})$$

$$\nabla f(\beta) = 2\Sigma_{\tilde{x}}\beta - 2\Sigma_{\tilde{y}\tilde{x}} = 0$$

$$\beta^* = \Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}}$$

Corresponding MSE

$$\begin{aligned}
&\mathrm{E}\left[(\tilde{y} - \tilde{x}^T \Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}})^2\right] \\
&= \mathrm{E}(\tilde{y}^2) + \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \mathrm{E}(\tilde{x}\tilde{x}^T)\Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}} - 2\mathrm{E}(\tilde{y}\tilde{x}^T)\Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}} \\
&= \mathrm{Var}(\tilde{y}) - \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1}\Sigma_{\tilde{x}\tilde{y}}
\end{aligned}$$

Additive model

Assume independent additive noise with zero mean $\tilde{y} = \tilde{x}^T \beta_{\text{true}} + \tilde{z}$

$$
\begin{aligned}
\mathrm{Var}(\tilde{y}) &= \mathrm{Var}(\tilde{x}^T \beta_{\text{true}} + \tilde{z}) \\
&= \beta_{\text{true}}^T \mathrm{E}(\tilde{x}\tilde{x}^T) \beta_{\text{true}} + \mathrm{Var}\left(\tilde{z}\right) \\
&= \beta_{\text{true}}^T \Sigma_{\tilde{x}} \beta_{\text{true}} + \mathrm{Var}\left(\tilde{z}\right)
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{\tilde{x}\tilde{y}} &= \mathrm{E}\left(\tilde{x}(\tilde{x}^T \beta_{\text{true}} + \tilde{z})\right) \\
&= \Sigma_{\tilde{x}} \beta_{\text{true}}
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{MSE} &= \mathrm{Var}(\tilde{y}) - \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} \\
&= \beta_{\text{true}}^T \Sigma_{\tilde{x}} \beta_{\text{true}} + \mathrm{Var}(\tilde{z}) - \beta_{\text{true}}^T \Sigma_{\tilde{x}} \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}} \beta_{\text{true}} \\
&= \mathrm{Var}(\tilde{z})
\end{aligned}
$$

What have we learned?

- Mean is best constant estimate in terms of MSE
- Conditional mean is best regression estimate in terms of MSE
  (but we often can't compute it)
- Best linear estimate only depends on covariance matrix of features,
  and covariance between features and response